Comparability Studies of the Virginia Computer-Delivered Tests

Steven Fitzpatrick
Pearson Educational Measurement

Robert Triscari
Virginia Department of Education

Paper Presented at AERA 2005

Abstract

The Commonwealth of Virginia began implementing its high school End-of-Course (EOC) tests in a computer administered mode in the fall of 2001. Since that time all 11 EOC tests have been implemented online. The original study in fall 2001 sought to establish comparability of three tests using a randomly equivalent groups design. After the data had been collected, it became apparent that the groups were not as equivalent as was hoped and a conceptual shift in the purpose of the study took place. Rather than attempt to establish comparability such that a single scoring table could be used for both modes of testing, it was decided that the online version of each test would be equated to its paper counterpart and each mode would have its own scoring table. A common item, non-equivalent groups design was used to accomplish this. In so doing, the performance standards would be "borrowed" from the paper test form the first time a subject was administered online. Subsequent test forms would be linked back to the previously administered online form. This paper presents the results of the analyses for the 11 EOC tests that are presently administered online. Finally, passing scores from the most recent administration of the EOC tests are presented as evidence that any effects due to mode of administration have been accounted for.

Introduction

As the Virginia Department of Education (VDOE) continues to move toward computerized delivery of the Standards of Learning (SOL) tests, research that examines the comparability of the computer-administered tests with the existing paper-and-pencil formats needs to be reported and replicated as additional SOL tests are included in the computerized delivery system. According to professional standards of practice, research studies should be conducted to demonstrate that students are not advantaged or disadvantaged in any way by taking the tests in an electronic form instead of the typical paper-and-pencil format. The American Psychological Association's *Guidelines for Computer-Based Tests and Interpretations* (1986) states: “when interpreting scores from the computerized versions of conventional tests, the equivalence of scores from computerized versions should be established and documented before using norms or cut scores obtained from conventional tests” (p. 18). Furthermore, the joint *Standards for Educational and Psychological Testing* (AERA, 1999) also recommends empirical validation of computerized versions of tests in Standard 4.11. “A clear rationale and supporting evidence should be provided for any claim that scores earned on different forms of a test may be used interchangeably” (p. 57).

Concern regarding a possible effect attributable to the computerization of a test stems from a number of possibilities. First, the human-computer interaction may interfere with test performance. It is important in any test to minimize the measurement errors associated with the student by test interaction and to ensure that such interaction does not interfere with what is being measured. Second, it is important that the medium of administration itself (i.e., the computer) does not introduce something unintended into the testing situation. The electronic test should measure student achievement in terms of the Standards of Learning; student achievement should not be determined by computer expertise. Associated with this are issues concerning how the computerization of a test may interfere with learned test-taking strategies. Finally, there may be issues related to whether item statistical information (i.e., item parameters) is directly transferable from paper-and-pencil mode to computerized mode. Ultimately, however, the main concern is the degree of equivalence between test scores from computerized and paper-and-pencil versions of the same test.

This paper describes the results of five comparability studies that have taken place since fall 2001. Table 1 lists the tests that have been studied and their corresponding administration dates. After the results from each study were examined, the decision was made as to whether to administer the tests studied online. For each of the tests studied thus far, the decision has been made to fully implement an online version of the assessment at the next administration. Although all of the studies thus far have shown that the results are comparable, it should also be noted, that due to the high stakes nature of the End-of-Course SOL tests, the online forms and paper forms are post equated separately. This ensures that even small mode effects are removed since different parameters and raw score to scaled score tables are generated for the paper and electronic versions of the same test. All of these tests contain only multiple-choice items. The length varies from 60 to 70 items with approximately 10 of the items being embedded field test items. All of the SOL tests are untimed. The only EOC test thus far not being delivered online is the English: Writing test. This test has both multiple-choice items and an extended response essay. The exploration as to when to bring this test online has begun.

Although at the outset it was the intention of these studies to show comparability, it was clear after the first study that in a high stakes graduation testing program, comparability in the sense that the same scoring table would be used for both modes was not the desired end result. Even small differences in scale scores at the proficiency cuts might result in different standards for paper and online delivery. Furthermore, in situations where the raw score to scale score tables were identical at the cut, there is no guarantee that future forms would not contain more mode effects than the form used for the study. Thus, separate equating studies have been and will be undertaken for the EOC tests. That is, after each administration, separate post equating analyses are conducted for the same form of the test administered online and on paper. Therefore, the conceptual shift that took place was to view the purpose of the studies as not to show comparability and thus combine or use the paper parameters for the online administration, but to view the purpose as, “can we borrow an equated cut score?”. That is, if the results of the study showed no more difference between the modes of administration than one would expect from alternate forms of the paper versions, we could link once to the paper scale and subsequently post equate each mode separately.

It should also be noted that the computer administration of the SOL tests is optional. If a student is not comfortable taking the test online, paper forms are available. Also, all of the “paper” accommodations such as Braille and large print test forms are still available for special education students. Now that these studies are complete, the Virginia Department of Education has begun to make some of the accommodations available on the paper tests, such as audio presentation of the items, available online.

The First Study: Fall 2001

Virginia uses an End-of-Course (EOC) model for its secondary school assessment program. Tests are administered in the fall, spring, and summer. Three different sets of base test items, referred to as Cores, are administered during each academic term. The primary form for a given term is called Core 1, and the secondary forms are called Cores 2 and 3. The Cores are rotated through the testing cycle with three new Cores developed each year.

Fifteen Divisions (School Districts) were invited to participate in the first comparability study. The study included three different EOC tests: Algebra I, Earth Science, and English: Reading, Literature, and Research. Students first took a Core 1 test in a paper administration. Students were then randomly assigned to a testing mode for the second administration and took either a paper or online version of a Core 2 test form. This test form had been administered as the primary paper test the prior spring. Students were told that the higher score from the live fall administration or the comparability study would be recorded as their EOC score in an attempt to maintain their motivation to perform well on the study test. Across the three subjects and two testing modes, 2205 students participated. There were from 292 to 463 students in each subject/mode combination.

The original intent of the fall 2001 study was to form two randomly equivalent groups of students with one group taking the paper version and the other taking the online version of the same test. Then the test could be calibrated in each mode and any differences could be attributed to mode effects. However, when the paper and online groups were compared using their scores on the primary administration of the Core 1 test, which all students took, the groups were found not to be as comparable as was hoped. This led to a revision of the analysis plans.

Since each of the Core 2 study test forms had been used for the primary paper administration in spring 2001, item parameter estimates and post-equated scoring tables were available for the paper mode. The decision was made to use a common item, non-equivalent groups design to link the results of the study data to the live paper scale. This is the same design that the state had been using to post-equate its paper tests across academic terms. Each test form has both forward and backward linking items. After a test is administered and calibrated, the mean of the backward links is set equal to the mean from the previous administration and the remaining items on the test are adjusted accordingly.

Table 2 shows the results of a Winsteps (1998) calibration of the Algebra I test in both online and paper modes. Items that have an '*' in the linking item column were treated as anchor items in the Winsteps run even though all of the items are common to both modes. This subset of items was used to anchor the calibration because they had been used to equate the test in the paper mode and we wanted the remaining items to be calibrated freely as this is where we would see any evidence of mode effects. The column labeled 'Paper Version Live Spring 2001' contains the post-equated item difficulties from the live administration. The remaining columns show the calibrated values for the same test administered in computer and paper modes during the study and the differences between them and those from the live administration. The values for linking items in the difference columns are displacements computed by WINSTEPS. The last few lines in the table summarize the differences between the three sets of item parameter estimates.

If mode effects were present for the online version of the test, they would be reflected in the displacements for the linking items or the difference between the live paper and online difficulty estimates for the non-linking items. When such differences do exist in the table, they are similar for the online and paper tests administered during the study. The average item difficulties from the three calibrations differ very little.

Table 3 shows the raw score to scale score tables resulting from the three calibrations of the Algebra I test. A scale score of 400 is the criteria for proficiency, and a score of 500 is considered advanced. The scale scores from the three calibrations differ by only a few points and in all cases the raw score cut for each performance category is the same.

Based on these results it appeared that there were no meaningful mode effects for the Algebra I test. Even so, rather than use the same raw score to scale score table for future paper and online forms, the decision was made that future online forms would be linked back to the adjusted item difficulties obtained from the online administration of this form.

The revision to the analysis plans because of the lack of equivalence of the groups assigned to the online and paper conditions led to a reconsideration of the purpose of the comparability studies for other EOC tests. Instead of trying to form randomly equivalent groups and conducting studies that attempted to establish the comparability of the two modes in the sense that a single scoring table could be used for both, it was decided to treat the issue as one of equating two tests using a common item, non-equivalent groups design. The online version of a test would be equated to its paper version and differences due to mode of administration would be accounted for in the resulting raw score to scale score table for the online version. In this way, each EOC test is linked to its paper version when the subject is first brought online and each subsequent online form is post-equated to the previous online administration.

Results of Online EOC Test Administrations

These studies examined and compared the results of live paper and computerized administrations of End-of-Course SOL tests. The test forms used in the comparability studies were always previously administered forms. These forms had item parameters and raw score to scale score tables from paper administrations. The data collection design was such that during a live administration of the test, students were given an alternate paper form of the test first and within several days they were given the online form of the test that had paper parameters. To motivate the students, they were given the higher of the two scores as their final result. The students were always allowed to take a paper form of the test first during the administration because of the high stakes nature (High School graduation) of the tests and the fact that the comparability in the online mode had not been established.

A paper-and-pencil test form for each subject area was converted to electronic images and used as the online form for the study. This conversion generated only minor modifications to a few items to better fit the computer presentation medium. Such as, items directing the student to "use the chart below" might have been reworded "using the chart above." Electronic testing specialists and psychometric staff reviewed the item changes and agreed that little or no threats to construct and content validity were introduced into the items by the conversion process.

The Rasch model was used to calibrate the tests using Winsteps. A common item non-equivalent groups design was used to place all of the parameters on the same scale. Although all of the items were common, only the items designated as links were used as anchors. During the computer based calibration runs, the linking items were anchored using the paper item parameters. Using this methodology, we were not able to differentiate the mode effect from other sources of differences in test difficulty. Since the intent was to "borrow" the equated cut score by post equating, it was not necessary to know how much of the change in difficulty was due to the mode effect and how much was attributable to the shift in test form difficulty. Thus, the one time link using the paper item parameters adjusts for both mode and form differences (since mode and form difficulty are really both measures of difficulty).

Each study utilized two different types of analyses to investigate comparability between the computer and paper-and-pencil testing conditions. First, item parameters estimates from the computerized tests were examined and compared with the item parameter estimates from the live paper-and-pencil administration. The differences between the item parameters for the non-linking items in the paper and online administrations were summarized as well as the displacements for the linking items. These displacements reflect the difference between the anchored value and what the calibration values would be if the items had not been fixed. During typical equating analyses, if these displacements exceed 0.50, the item is further examined and may be dropped from the linking set. The second analysis was a comparison of the raw score to scale score tables in light of the item parameter differences.

The results from all of the studies are summarized in Table 4. The raw score cuts at the proficiency level remained the same for the online test as they were on the paper tests for Algebra I, Algebra II, Biology, Chemistry, and Geometry. The average difference between the paper and online modes for the non-linking items ranged from -0.089 to 0.032 for these tests. The raw score cuts for the Earth Science, English: Reading, Literature, and Research, and World Geography tests were one point lower in the online mode than in the paper mode. The differences between the non-linking items parameters in the two modes for these tests ranged from 0.099 to 0.157 indicating that the online version was slightly more difficult. This is accounted for in the lower cut score for the online version. In contrast, the raw score cuts for the three history tests increased. The differences between the non-linking item parameters in the two modes for these tests ranged from -0.087 to -0.172 suggesting that the online version was slightly easier.

Table 5 shows the post-equated results from the live spring 2004 administration. Approximately equal numbers of students took the online and paper versions of the test in each subject area. Several points should be noted. First, for the 20 paper to online comparisons, 11 of the cut scores were the same, four differed by one raw score, four differed by two, and one differed by three (Earth Science – Core 2). Second, for the ten paper form comparisons (Paper Core 1 to Paper Core 2), two of the raw score cuts were the same, six raw score cuts differed by one, one differed by two, and one differed by three. And, third, this table shows that no discernable differential drift seems to be taking place across modes. That is, the differences across modes are less than or equal to the results found in the earlier studies. These data help to show that equating once across modes with the intent of "borrowing" the equated cut scores results in equating differences that are no greater than within paper test form equating differences.

References

American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME) (1999*), Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

American Psychological Association (1986*). Guidelines for Computer-Based Tests and Interpretations*, Washington, DC: American Psychological Association.

Linacre, M. J., (1998). WINSTEPS. Chicago, IL: Mesa Press.

Table 1.

Comparability Studies of Paper and Pencil and Online SOL Assessments

Testing Session	**SOL Assessments Studied**
Fall, 2001	Algebra I, Earth Science, English: Reading
Spring 2002	Algebra II, Biology
Fall, 2002	VA & US History, World History I, World History II
Spring, 2003	World Geography, Chemistry
Spring 2004	Geometry

Table 2
Algebra I -- Fall 2001
Items

Item Number	UIN	Linking Item	Paper Version Live Spring 2001	Computer Version Anchored to linking items	Computer minus Spring 2001 Delta*	Paper Version Anchored to linking items	Paper minus Spring 2001 Delta*
1			0.035	-0.262	-0.297	-0.387	-0.422
2		*	-1.229	-1.229	0.020	-1.229	-0.030
3			0.978	0.821	-0.157	0.755	-0.223
4			-0.734	-0.274	0.460	-0.156	0.578
5			-0.758	-1.273	-0.515	-0.963	-0.205
6		*	-0.507	-0.507	0.070	-0.507	-0.110
7			0.305	-0.007	-0.312	-0.156	-0.461
8		*	-0.023	-0.023	0.050	-0.023	-0.170
9			-0.609	-1.290	-0.681	-0.857	-0.248
10		*	-0.278	-0.278	-0.250	-0.278	0.070
11			-0.073	-0.394	-0.321	-0.245	-0.172
12		*	0.152	0.152	0.140	0.152	0.090
13			-0.291	-0.346	-0.055	-0.296	-0.005
14			1.241	1.134	-0.107	1.133	-0.108
15			0.208	0.152	-0.056	0.143	-0.065
16		*	0.795	0.795	-0.180	0.795	-0.050
17			0.487	0.413	-0.074	0.610	0.123
18			0.688	0.311	-0.378	0.268	-0.420
22		*	0.504	0.504	0.120	0.504	0.530
23			1.058	1.057	-0.001	0.988	-0.070
24			0.251	0.493	0.242	0.469	0.218
25			1.186	1.108	-0.078	0.974	-0.212
26			0.060	-0.169	-0.229	-0.322	-0.382
27			1.372	1.161	-0.211	1.045	-0.327
28			-0.844	-0.737	0.107	-0.813	0.031
29		*	-0.092	-0.092	0.140	-0.092	0.340
30			-0.137	0.379	0.516	0.482	0.619
31		*	0.174	0.174	0.090	0.174	0.050
32			-0.418	-0.134	0.284	-0.271	0.147
33		*	-0.862	-0.862	0.140	-0.862	0.220
38		*	-0.355	-0.355	-0.030	-0.355	0.110
39			0.342	-0.030	-0.372	0.131	-0.211
40			0.318	0.504	0.186	0.571	0.253
41		*	1.407	1.407	0.080	1.407	-0.130
42			0.762	0.906	0.144	0.597	-0.165
43			0.516	0.631	0.115	0.218	-0.298
44			0.494	0.797	0.303	0.494	0.000
45			-0.590	-0.917	-0.327	-0.479	0.111
46		*	0.190	0.190	-0.280	0.190	-0.230
47			1.216	1.018	-0.198	1.017	-0.200

48		*	0.126	0.126	-0.340	0.126	-0.370
49			1.349	1.424	0.075	1.286	-0.063
53		*	0.346	0.346	-0.320	0.346	-0.450
54			0.177	0.243	0.066	-0.194	-0.371
55			-0.729	-0.671	0.058	-0.670	0.059
56			-1.550	-1.969	-0.419	-1.718	-0.168
57			-0.632	-0.975	-0.343	-0.813	-0.181
58		*	0.511	0.511	0.170	0.511	0.230
59		*	0.146	0.146	0.270	0.146	-0.130
60			0.369	0.631	0.262	0.181	-0.188
Mean			0.141	0.095	-0.048	0.081	-0.061
Difference in Means					-0.046		-0.061
Root mean squared difference					0.259		0.260
Average absolute difference					0.213		0.212

*Note: The values for linking items in the difference columns are displacements computed by WINSTEPS.

Table 3
Algebra I Fall 2001
Scale Scores

		Computer Version		Paper Version	
Raw Score	Live Spring 2001 Adjusted	Anchored to Spring 2001	Computer minus Spring 2001	Anchored to Spring 2001	Paper minus Spring 2001
1	216	211	-5	213	-3
2	248	243	-5	245	-3
3	267	263	-4	265	-2
4	281	277	-4	279	-2
5	293	289	-4	291	-2
6	302	299	-3	300	-2
7	311	307	-4	308	-3
8	318	315	-3	316	-2
9	325	322	-3	323	-2
10	331	328	-3	329	-2
11	337	334	-3	335	-2
12	342	340	-2	340	-2
13	347	345	-2	345	-2
14	352	350	-2	350	-2
15	357	355	-2	355	-2
16	362	359	-3	359	-3
17	366	364	-2	364	-2
18	370	368	-2	368	-2
19	375	373	-2	372	-3
20	379	377	-2	376	-3
21	383	381	-2	380	-3
22	387	385	-2	384	-3
23	391	389	-2	388	-3
24	395	393	-2	392	-3
25	399	397	-2	396	-3
26	403	401	-2	400	-3
27	407	405	-2	404	-3
28	411	409	-2	408	-3
29	415	413	-2	412	-3
30	419	417	-2	416	-3
31	423	422	-1	420	-3
32	427	426	-1	424	-3
33	431	430	-1	429	-2
34	436	435	-1	433	-3
35	440	439	-1	438	-2
36	445	444	-1	442	-3
37	450	449	-1	447	-3
38	455	454	-1	452	-3
39	461	459	-2	458	-3
40	466	465	-1	463	-3
41	473	471	-2	469	-4

Table 3
Algebra I Fall 2001
Scale Scores

		Computer Version		Paper Version	
Raw Score	Live Spring 2001 Adjusted	Anchored to Spring 2001	Computer minus Spring 2001	Anchored to Spring 2001	Paper minus Spring 2001
42	479	478	-1	476	-3
43	487	486	-1	483	-4
44	495	494	-1	492	-3
45	504	503	-1	501	-3
46	516	515	-1	512	-4
47	530	529	-1	527	-3
48	549	548	-1	546	-3
49	581	580	-1	578	-3

Table 4
Average Displacements and Differences Between Paper and Online EOC Test Administrations

Test	Number of Items	Number of Links	Sample Size	Average Displacement	Average Difference	Paper Raw Score Cut	Online Raw Score Cut
Algebra I	50	17	395	-0.006	-0.070	26	26
Earth Science	50	14	463	-0.012	0.099	29	28
English:RLR	42	9	300	-0.016	0.157	22	21
Algebra II	50	16	1287	-0.009	0.032	31	31
Biology	50	12	1878	0.006	-0.015	27	27
VA and US History	61	13	1334	-0.007	-0.087	37	38
World History I	61	14	1734	-0.010	-0.115	32	34
World History II	61	16	1345	-0.016	-0.172	31	33
World Geography	60	17	3512	-0.008	0.132	29	28
Chemistry	50	17	3698	0.001	-0.014	27	27
Geometry	45	14	3451	0.011	-0.089	27	27

Table 5
Spring 2004 Live SOL Cut Scores for Online and Paper Tests

Test	Core 1		Core 2	
	Online	Paper	Online	Paper
Algebra I	28	29	27	27
Earth Science	30	32	29	32
English: RLR	27	25	27	26
Algebra II	29	31	30	30
Biology	28	28	28	28
VA and US History	31	31	28	28
World History I	31	31	30	30
World History II	28	30	29	29
World Geography	32	33	32	32
Chemistry	26	27	26	26